European Journal of Operational Research 219 (2012) 611–621

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming

Verena Schmid

Department of Business Administration, University of Vienna, Bruenner Strasse 72, 1210 Vienna, Austria

ARTICLE INFO

Article history:
Available online 10 November 2011

Keywords:
OR in health services
Emergency vehicles
Ambulance location
Approximate dynamic programming
Stochastic optimization

ABSTRACT

Emergency service providers are supposed to locate ambulances such that in case of emergency patients can be reached in a time-efficient manner. Two fundamental decisions and choices need to be made real-time. First of all immediately after a request emerges an appropriate vehicle needs to be dispatched and send to the requests' site. After having served a request the vehicle needs to be relocated to its next waiting location. We are going to propose a model and solve the underlying optimization problem using approximate dynamic programming (ADP), an emerging and powerful tool for solving stochastic and dynamic problems typically arising in the field of operations research. Empirical tests based on real data from the city of Vienna indicate that by deviating from the classical dispatching rules the average response time can be decreased from 4.60 to 4.01 minutes, which corresponds to an improvement of 12.89%. Furthermore we are going to show that it is essential to consider time-dependent information such as travel times and changes with respect to the request volume explicitly. Ignoring the current time and its consequences thereafter during the stage of modeling and optimization leads to suboptimal decisions.

1. Introduction and related work

Emergency service providers are supposed to locate ambulances such that in case of emergency patients can be reached in a time-efficient manner. Two fundamental decisions and choices need to be made real-time. First of all immediately after a request emerges an appropriate vehicle needs to be *dispatched* and send to the requests' site. Ambulances, when idle, are located at designated waiting sites. Hence after having served a request the vehicle needs to be *relocated* (i.e. its next waiting site has to be chosen). For a close match to reality, time-dependent information for both traveling times and the request volume will be considered explicitly. We are going to solve the underlying optimization problem using approximate dynamic programming (ADP), an emerging and powerful tool for solving stochastic and dynamic problems typically arising in the field of operations research.

In practice simple rules for dispatching and relocation are in use. In Austria, because of regulatory rules, in case of an emergency always the closest ambulance will be dispatched. After having served a request ambulances are supposed to return to their home base. Using ADP we are going to relax these assumptions and propose different strategies in order to improve the performance of the underlying system and its capability to efficiently serve emergency requests. Empirical studies suggest that after for instance a cardiac and circulatory arrest the chances for a resuscitation to be successful decrease dramatically. Typically chances decrease by 10% per minute as long as the patient is not treated accordingly. Providing a quick response to emergency requests is crucial for the patients' state of health.

The contribution of this paper is threefold.

(i) We propose a stochastic dynamic model for the ambulance relocation and dispatching problem, which will be solved by means of ADP.
(ii) In order to get a preferably accurate model of reality we will explicitly take into account time-dependent information and variations with respect to changing request volumes and travel times, varying throughout the course of the day.
(iii) We are able to improve dispatching and relocation strategies currently in use.

Different possibilities exist for measuring and evaluating the performance of emergency service providers. Typically timeliness is considered as the primary objective. Most performance measures are related to response or waiting times. In this paper we will try to minimize the average response time observed. We can show that it is not necessarily a good choice to always dispatch the closest vehicle available. From a global perspective it does make sense (whenever the level/priority of the request permits) to send a vehicle that might be slightly farther away. As a consequence the waiting time observed by this particular request will be marginally

E-mail address: verena.schmid@univie.ac.at

0377-2217 © 2011 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.ejor.2011.10.043

worse. From a global perspective however the remaining vehicles are located more effectively in order to cover future demands more efficiently. By making decisions in an anticipatory manner, while implicitly taking into account the current situation as well as potential future requests, the performance of the emergency service provider can be improved. The algorithm has been tested on real data form the city of Vienna and has been benchmarked against two myopic policies traditionally in use.

We are going to show that it is essential to consider time-dependent information such as travel times and changes with respect to the request volume explicitly. Ignoring the current time and its consequences thereafter during the stage of modeling and optimization leads to suboptimal decisions. Furthermore we have done extensive parameter testing for the proposed algorithm in use.

In the last decades several location models for the health care sector have been discussed. Typically these include the optimization of the location of ambulance vehicles, such that the population can be covered (i.e. reached) in case of an emergency effectively. Most models found in the literature are extensions of the location set covering model by Toregas et al. (1971) or the maximal covering location model by Church and ReVelle (1974). Whereas the former which tries to minimize the number of ambulances in use and the latter tries to optimize the demand covered using a fixed fleet size. These are static location problems that do not consider the fact that resources might become unavailable over time. Also the routing aspect itself is not yet present. One possibility for tackling the former concern includes multiple coverage, i.e. demand points are supposed to be covered by more than one vehicle. Such a model, called the double standard model (DSM) was introduced by Gendreau et al. (1997). Their model has been further extended with respect to capacity considerations by Doerner et al. (2005) and successfully applied to eight rural provinces in Austria. Other successful applications of models based on DSM formulations include Thirion (2006) for Belgium and Gendreau et al. (2001, 2006, 1997), who applied their models to data from Montreal.

Other possibilities for handling vehicles becoming unavailable include busy functions, where the probability that vehicles might become unavailable is modeled explicitly (see Daskin, 1983).

Several approaches already have been proposed that handle the ambulance location problem in a dynamic setting. An extension of the maximum expected coverage location problem proposed by Daskin (1983) has been developed by Repede and Bernardo (1994). Within their paper the authors propose and solve a multi-period maximum expected coverage location problems with time-variant travel times and changing fleet sizes. A similar approach, also taking into account the resulting number of relocations, has been developed by Schmid and Doerner (2010). A dynamic model for real-time ambulance relocation has been proposed in Gendreau et al. (2001), where different redeployment scenarios are precomputed. Another multi-period model for dynamic demand environment which minimized the number of ambulances required while meeting predetermined ambulance availability requirements has been proposed in Rajagopalan et al. (2008). A broad overview on different location problems and their applications in the context of ambulance location problems can be found in Brotcorne et al. (2003), Laporte et al. (2009).

We will solve the problem at hand using ADP, a very powerful approach for modeling and solving stochastic and dynamic optimization problems. Decisions (such as the choice *which* vehicle should be dispatched in case of a request, or *where* the vehicle should be sent afterwards, etc.) have to be made over time under uncertainty. Decisions that have to be made now do have an impact on the future; hence we need to find a way to anticipate their effect on future consequences (i.e. the capability of the system to serve future requests). Due to the well-known curses of dimensionality most large-scale stochastic optimization problems cannot be solved to optimality.

For a general discussion of different algorithms related to the concept of ADP we refer to Powell (2007). Similar concepts exist in other communities such as control theory (see Bertsekas and Tsitsiklis, 1996 for an overview on *neuro-dynamic programming*) and the computer science and artificial intelligence community (see Sutton and Barto, 1998 for an overview on *reinforcement learning*).

ADP has been applied successfully to resource allocation problems (see Powell et al., 2001; Godfrey and Powell, 2002a,b) and large-scale fleet management (see Simão et al., 2009; Powell and Topaloglu, 2005). For making ambulance redeployment decisions in a dynamic setting under uncertainty an ADP approach based on a policy iteration algorithm has been developed by Maxwell et al. (2009, 2010). In their papers however the authors focus on redeployment decisions (for *idle* vehicles) only, while maximizing the number of calls reached within a given delay threshold. In Austria however, by law, repositioning idle ambulances (apart from sending them back to a waiting location) are not allowed. Hence we try to compensate for vehicles currently being busy and the systems' capability to cover future requests by different dispatching and relocation strategies. Furthermore, in comparison to our model, they assume travel times and call arrival rates to be constant over time.

We will give a detailed description of the model in Section 2. A mathematical formulation for the underlying dynamic and stochastic problem will be provided in Section 3. Our solution approach, which is based on ADP, will be presented in Section 4. The obtained solutions and results, as well as some additional insights and evaluations, will be provided in Section 5.

2. Problem description

Requests for emergency transportation only become known at very short notice. It is important that the system is highly flexible and robust in a sense that allows to quickly sending vehicles to the emergency site in cases needed. Hence it is crucial that idle vehicles are located and dispersed throughout the geographic area under consideration such that emergency patients can be reached quickly. The response time (i.e. the time necessary from the arrival of the call until the vehicle finally reaches the corresponding location) is a common quality characteristic for measuring the performance of ambulance dispatching services.

The dispatching process itself can be described as follows. See Fig. 1 for a graphical representation of the dispatching process.

A *request* typically arrives by phone and is answered by a dispatcher who enters all relevant data into the dispatching system and, using a predefined set of questions, determines the priority of the call. The time at which the emergency request r becomes known to the system is denoted by t_r. In case a suitable idle vehicle is available it will be assigned and it is supposed to set off towards the corresponding patient location right away (at time a_r). The total dispatching time required (i.e. the time necessary from the arrival of the request until a vehicle can be assigned) is denoted by dt_r. This time span typically includes the time necessary for inquiring information concerning the actual incident, identifying an adequate ambulance and typically a setup time required for the crew to get ready. Ambulances arrive (after driving for tt_r^p time units) at the call's scene and start their first-aid measures at time s_r^p. Service is completed (after st_r^p time units) and the corresponding ambulance leaves the call's site at e_r^p. The ambulance reaches the final destination (typically a hospital) after additional tt_r^h units at s_r^h, where the crew starts to unload the patient and she will be admit into the corresponding department. We assume that there are no setup-times required between individual events. For instance that